

2017 6.12-6.18 周报

Junhua Lu

Done

1. Extract wiki data under 'movie', including edit history and discussion records. A preliminary analysis on different kinds of users (bot, ip as username, registered users) are conducted, including their respective contributions to page(article) and talk, and the corresponding distributions.

wikidata

Saturday, June 17, 2017 9:18 PM

不知道性别196,156人

总共1,153,011人

机器人也是既有talk 也有article的

机器人一共162个, (要用rights而不是groups来判断)

贡献talk总数42,951, 去重后24,318

贡献article总数172,775, 去重后37,313

IP地址相关

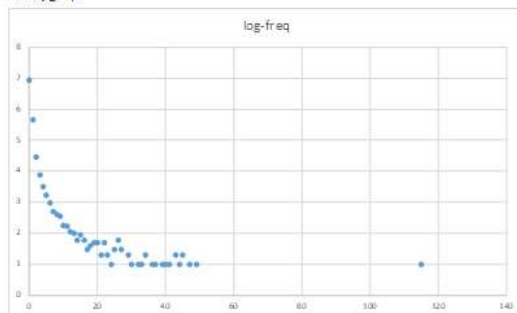
统计下ip地址发帖总数, 以及发帖分布

930,037人是ip发帖的, 占了80.66%

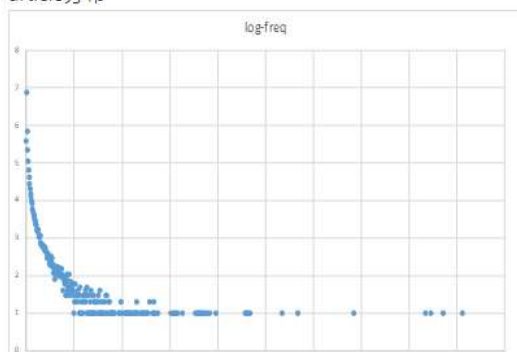
发talk总数61,245 去重后11,434

发article总数1,351,165 去重后37,106

talk分布



article分布



除此以外的用户的统计情况

暂时叫做normal user

一共220,231人

一共talk 296,367 去重后41,292

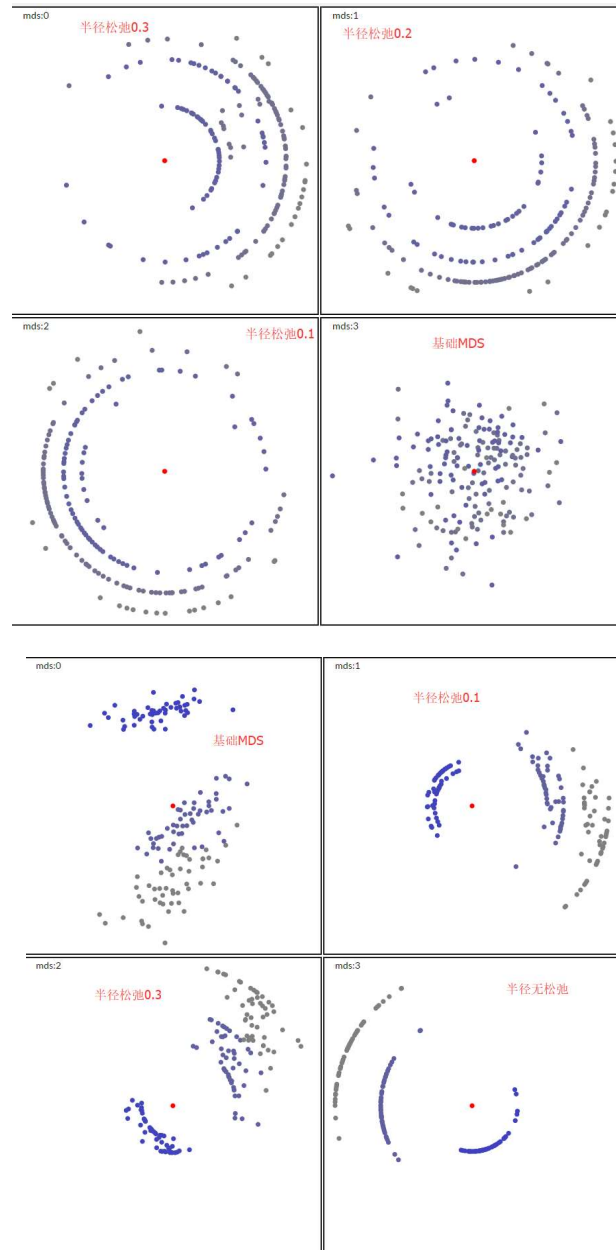
一共article1,710,354 去重后41,718

在统计频次时候, 都发现article中频次为1的多于为0的

Talk 41,771

Article 当然也是这么多

2. A discussion on radial projection and revision on this year's submission. The projection is not bad if applied to existing well-classified datasets, and I will use real dataset of players to examine its performance.



3. Listened to a talk on ‘Turning Massive Text Corpora into Structures’, which offered us a different way to machine-understand documents, like entity typing, which seems like ontology or so. From my point of view, it is different from the NLP in my understanding.
4. A abstraction and minor revisions are added to 专著.

To Do

1. Revise minor issues on BotRadar
2. Pre-processing on wiki data, based on future discussions

Papers

VAST 16 Vast Best Paper An Analysis of Machine- and Human-Analytics in Classification 这是一篇从信息论角度分析在分类问题中可视化辅助人带来的优势的分析, 里面对一般文章不度量的 soft knowledge 也尝试了度量.

KDD16 CatchTartan: Representing and Summarizing Dynamic Multicontextual Behaviors.

也是之前 Meng Jiang 做的, 之前他们的作品是检测 dense block, 类似下面这个工作, 现在又说这些张量方法其实不够好(无法好好体现 dynamic 或者 evolution), 而用了这种有点可视化成分的方法, 先将随时间变化的东西用二维形式表示出来, 再用一个类似于压缩的东西(用到了表示距离)来找出其呈现的 Tartan(苏格兰裙格), 其实也是在寻找一种特殊的事件块.

KDD 16 Fraudar: bounding graph fraud in the face of camouflage

问题从 fake review 出发, 类似于之前 dense block 的问题. 一般来说会有有用谱聚类 and 置信传播(belief propagation)的方法 现在有一些假的 review, 用 camouflage 躲避检查, 通过 adding reviews or follows with honest targets so that they look “normal”. Hijack 其他账户来添加 review. 问题定义: 二部图(users/products) (optional: 先验节点嫌疑度), 定义一个函数 $g(A, B)$ 返回标量值, A 代表 user B 代表 products 这样一块. 所谓 camouflage-resistant, 就是 $g(A, B)$ 不会下降如果 camouflage 加入 A 的话. 定义了一个非常简单的度量 $= (\text{节点危险度之和} + \text{边危险度之和}) / (|A| + |B|)$. 这里有个边危险度, $C_{ij} = 1 / \log(\text{第 } j \text{ 列的非加权和})$. 有时候密集块可能是商品太畅销导致比如咖啡. Fraudlent 商品有很高的比例的边来自 fraud. Log 源自 tf-idf 思想, 最后贪婪算法从整个图开始不断删除可能的行或者列, 删了就要评估下剩下的图, 一直一直试过去. 采取一定加速方法可以降到线性复杂度.

KDD16 Come-and-Go Patterns of Group Evolution: A Dynamic Mode

社交群, 比如微信群什么的加群退群是怎样的模式, 有没有一个可以概述所有的模型呢? 从数据发现问题, 考察了 103548 个微信群, 经典的 SIR(传染病模型)和 SI 模型 还有什么幂律 decay 都只能解释部分问题. 本文提出的模型先从基本的 SI 和 SIR 入手, 由 SIR 导出群内人员在群停留时间的分布, 作为一个模型的一个基础知识. 对整个模型, $I(t)$ 表示表示 t 时群里人, $J(t)$ 表示在 t 时加入的累积人数, $Q(t)$ 表示 t 时前累积退出的. 进入的仍然用旧的 word-of-mouth 模型, 退出则是用上面说的停留分布来导出. 在此基础上, 还考虑非扩散因素导致的群增长, 这里举的例子是一个新建的社团除了靠朋友拉人, 还有可能举办活动来拉人, 这样会产生一个 external shock 类似的造成的 burst. 加入以后模型一共有六个参数, 每个参数都有合理的物理含义. 虽然发在 KDD, 但似乎对社会科学方面的贡献可能会更大些, 个人认为.